Using Lake House approach to enable self-service analytics with AWS Lake Formation

Roy Hasson – Pr. Product Manager

Customers want more value from their data

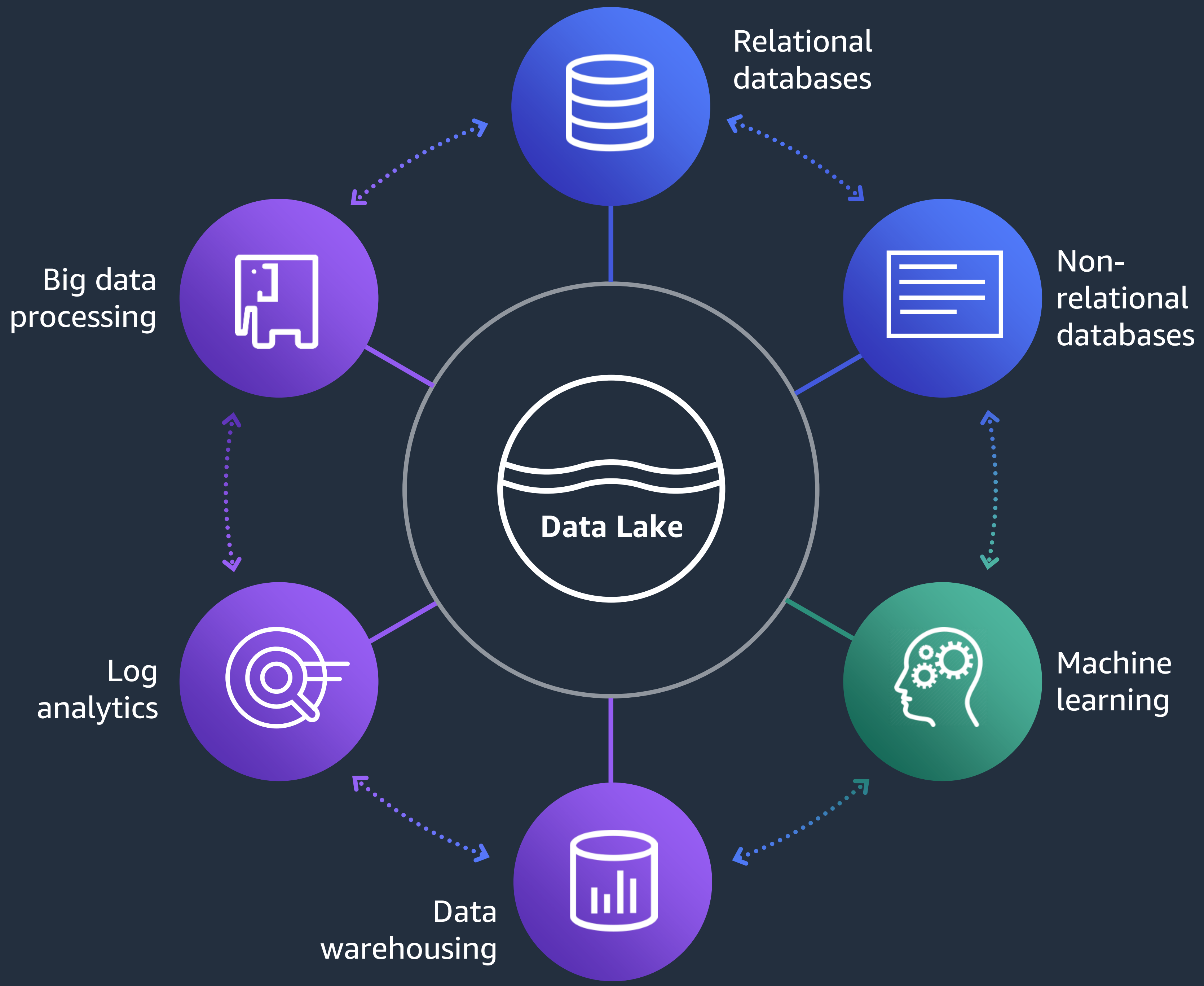
GROWING EXPONENTIALLY

FROM NEW SOURCES

INCREASINGLY DIVERSE

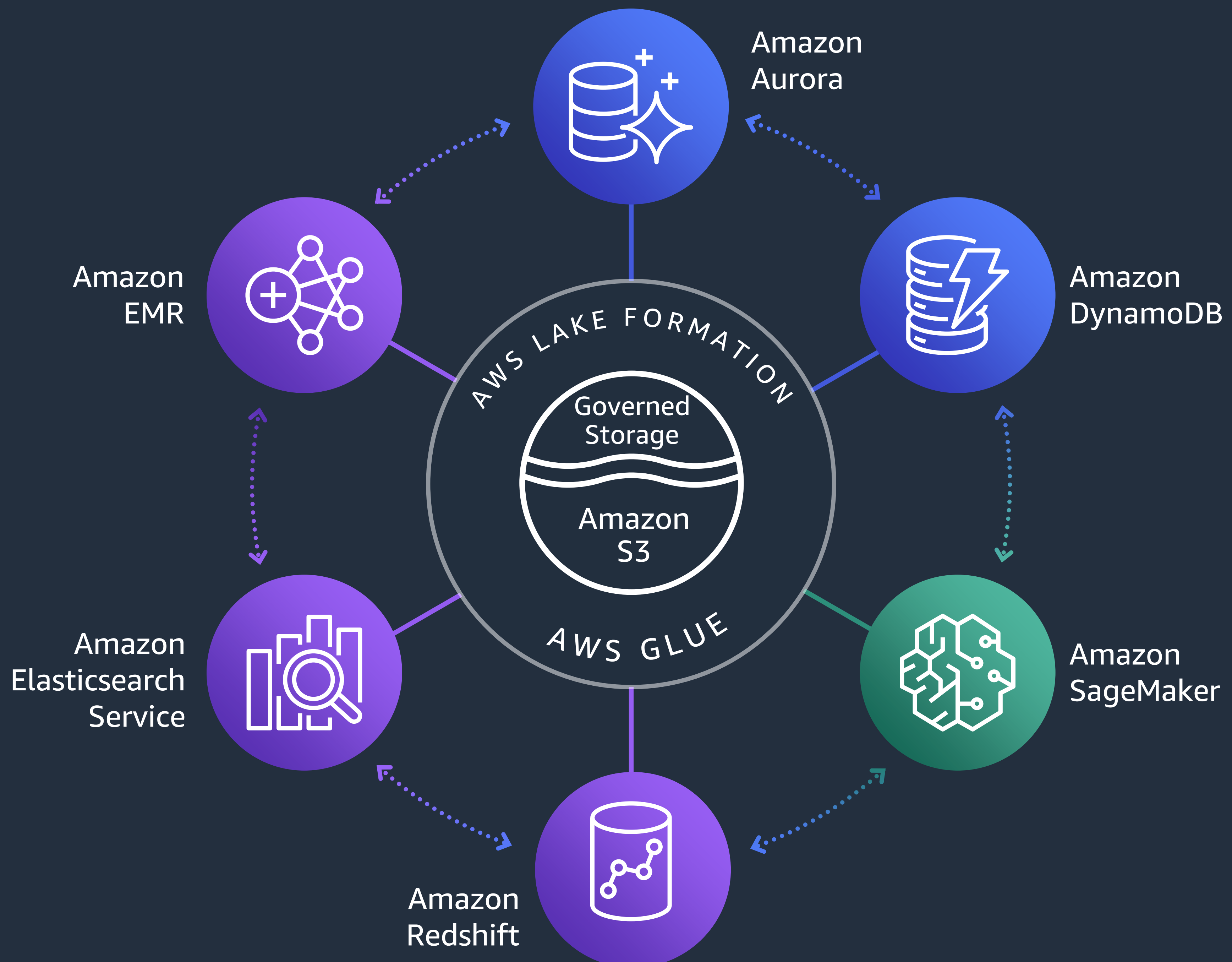
USED BY MANY PEOPLE

ANALYZED BY MANY APPLICATIONS



Customer data in the real world

Lake House approach on AWS



SCALABLE DATA LAKES

PURPOSE-BUILT DATA SERVICES

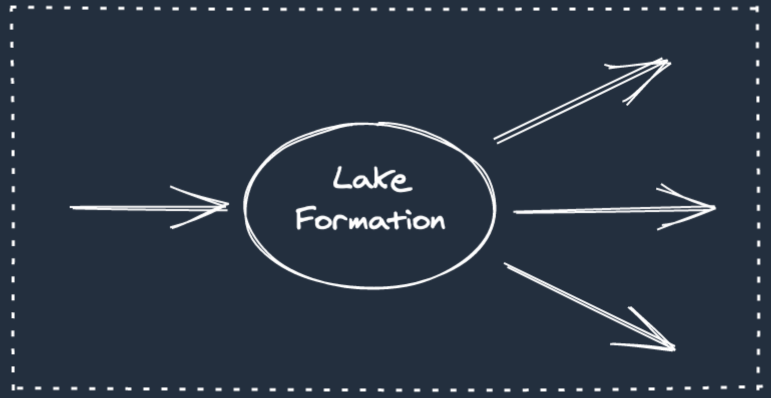
AUTOMATED DATA MOVEMENT

CENTRAL GOVERNANCE

PERFORMANT AND COST-EFFECTIVE

Common architecture design patterns

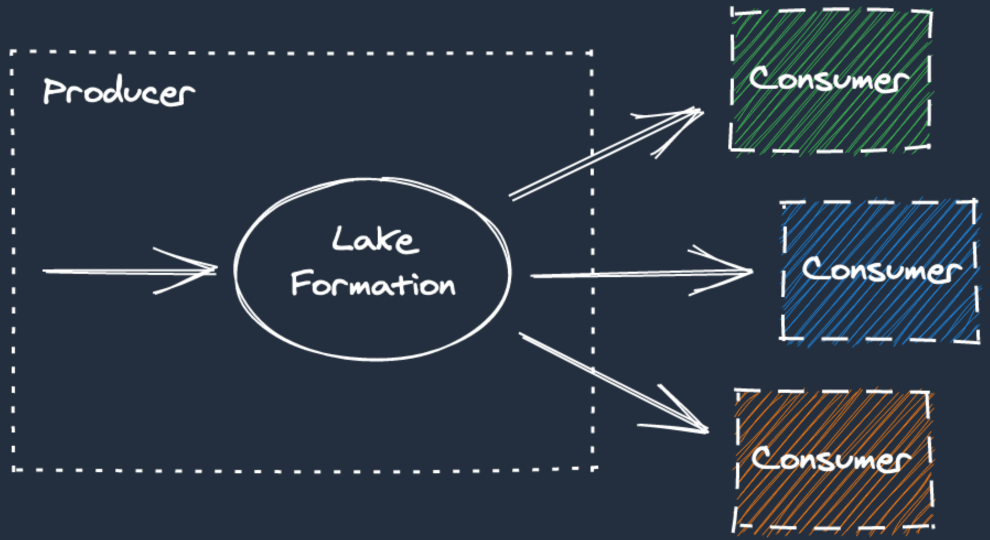
SINGLE ACCOUNT



CENTRALIZED
SINGLE ACCOUNT

SIMPLE TO GET STARTED

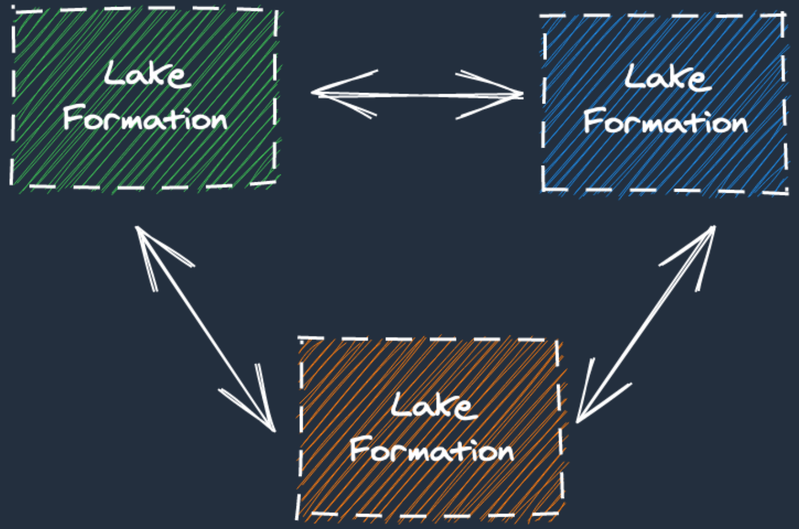
HUB & SPOKE



HUB & SPOKE
MULTI-ACCOUNT

CROSS ORGANIZATION

DATA MESH



DATA MESH
CENTRAL GOVERNANCE

ORGANIZATIONAL AUTONOMY

Amazon S3 is the most popular storage for data lakes

Unmatched **durability, availability,** and **scalability**

Easiest to use with **cost optimization:** Intelligent tiering

Most ways to **get data in**

Most **object-level controls**

Broadest portfolio of analytics tools

Best **security, compliance,** and **audit** capabilities

Cold storage and archive capabilities

Operationalizing data lakes on S3 is still hard

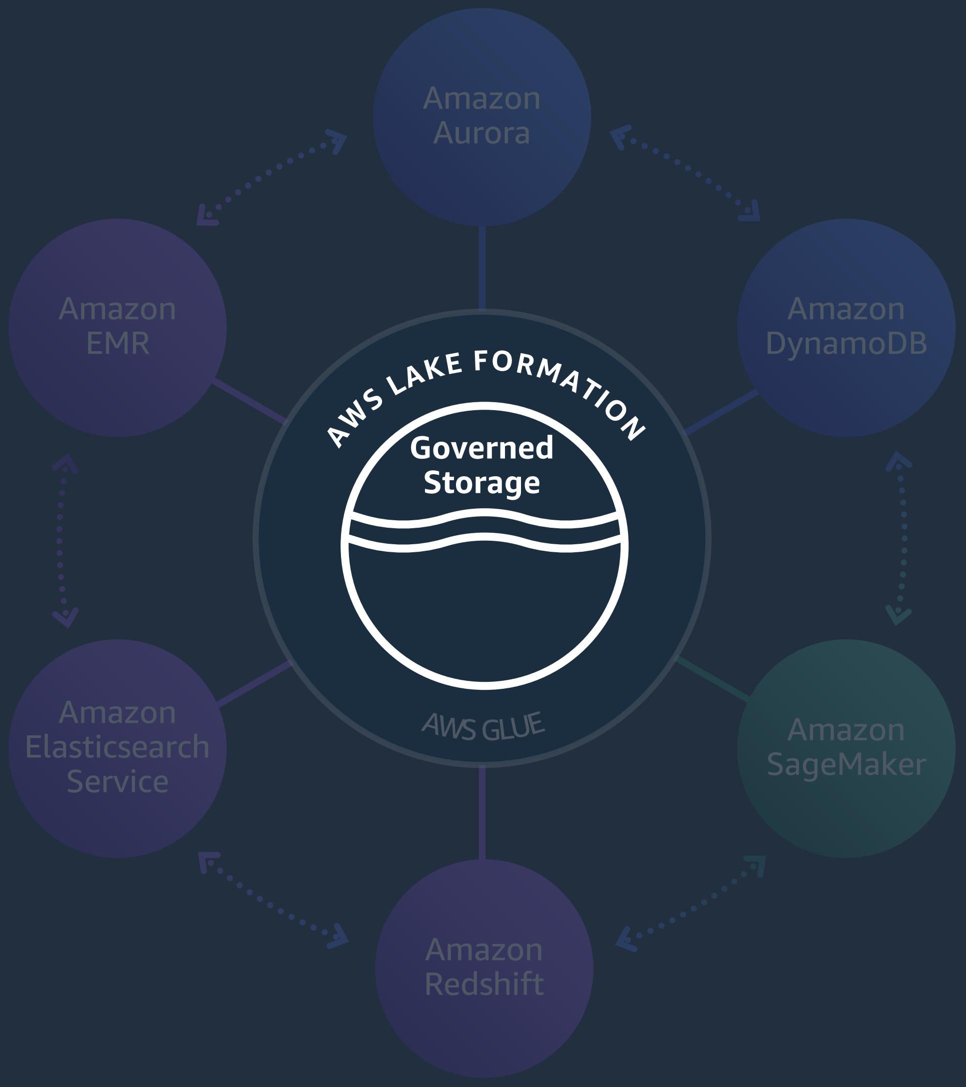
DIFFICULT TO INSERT, UPDATE AND DELETE DATA

SECURITY AND GOVERNANCE IS COMPLEX

ERROR-PRONE COARSE GRAIN DATA SHARING

AWS Lake Formation

Build a secure data lake in days



Build data lakes quickly

Move, store, update, and catalog your data faster
Automatically organize and optimize your data

Simplify security management

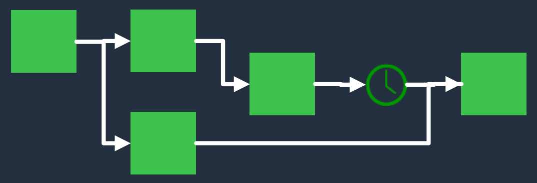
Centrally define and enforce security, governance, and auditing policies

Easily discover and share data

Catalog all of your data assets and easily share datasets between consumers

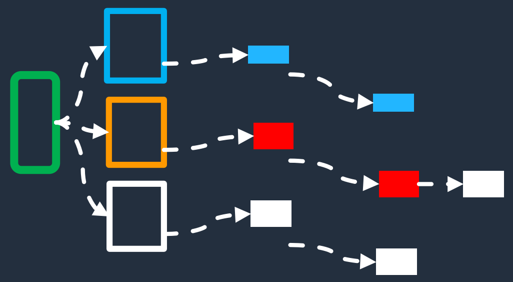
Challenges in managing your data

CONTINUOUS UPDATES



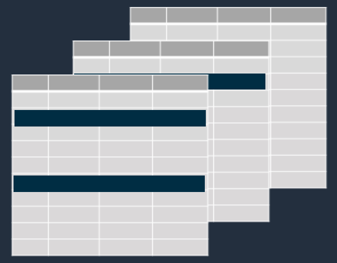
COMPLEX ETL
DELAYS IN DATA FRESHNESS
EXPENSIVE, BRITTLE & ERROR-PRONE

INCONSISTENT PERFORMANCE



OVER-SCAN DATA
LOTS OF SMALL FILES
PARTITION UPDATES
MANAGEMENT OVERHEAD

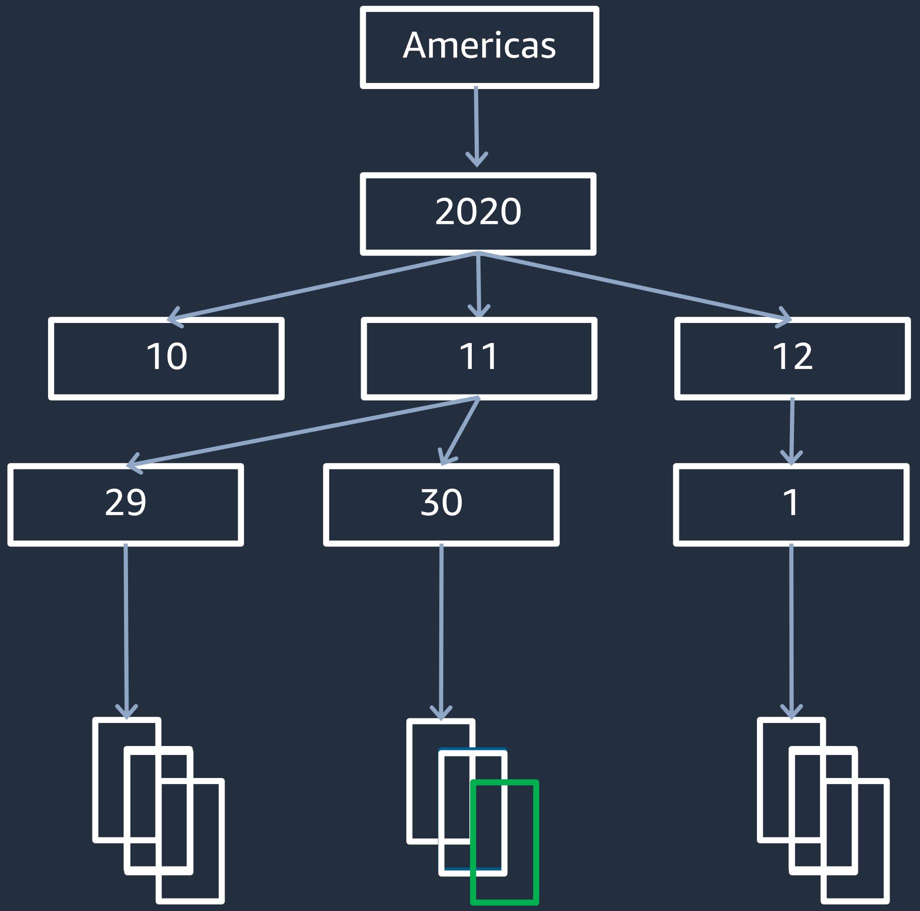
COMPLYING WITH REGULATIONS



DIFFICULT TO FIND NEEDLE IN VERY LARGE HAYSTACK

Standard external tables on S3

Apache Hive-style tables, organize data in *partitions*



S3 path following Hive-style partition

Files in open formats

S3://IOTDeviceData/region=Americas/year=2018/month=Nov/day=30/data1.parquet
S3://IOTDeviceData/region=Americas/year=2018/month=Nov/day=30/data2. parquet
S3://IOTDeviceData/region=Americas/year=2018/month=Nov/day=30/data3. parquet
…

Introducing Governed Tables

PREVIEW

New type of S3 table

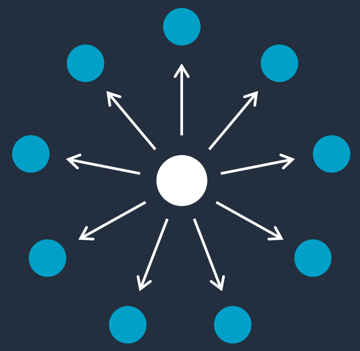
UPDATE DATA

ACID TRANSACTIONS

CONSISTENT ACROSS TASKS
INSERT, UPDATE, DELETE
CONVERGE BATCH & REAL-TIME

RELIABLE

OPTIMIZATION ACCELERATION



STORAGE OPTIMIZATION

AUTO-COMPACT SMALL FILES
PUSH-DOWN FILTERS
REDUCE DATA SCAN

PERFORMANT

TIME TRAVEL

DATA VERSIONS

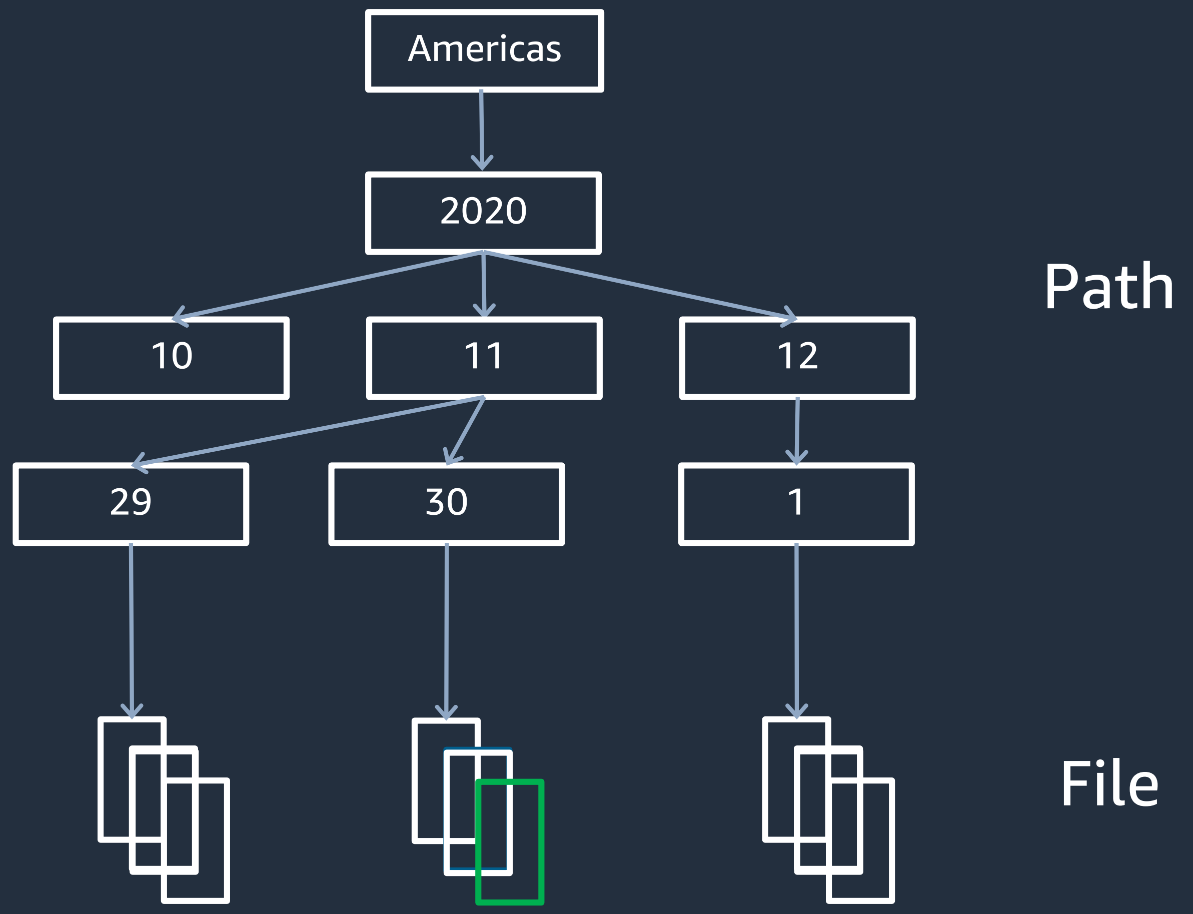
DATA HISTORY
REPRODUCE EXPERIMENTS
AUDIT CHANGED DATA

VERSIONED

Governed Tables: Under the hood

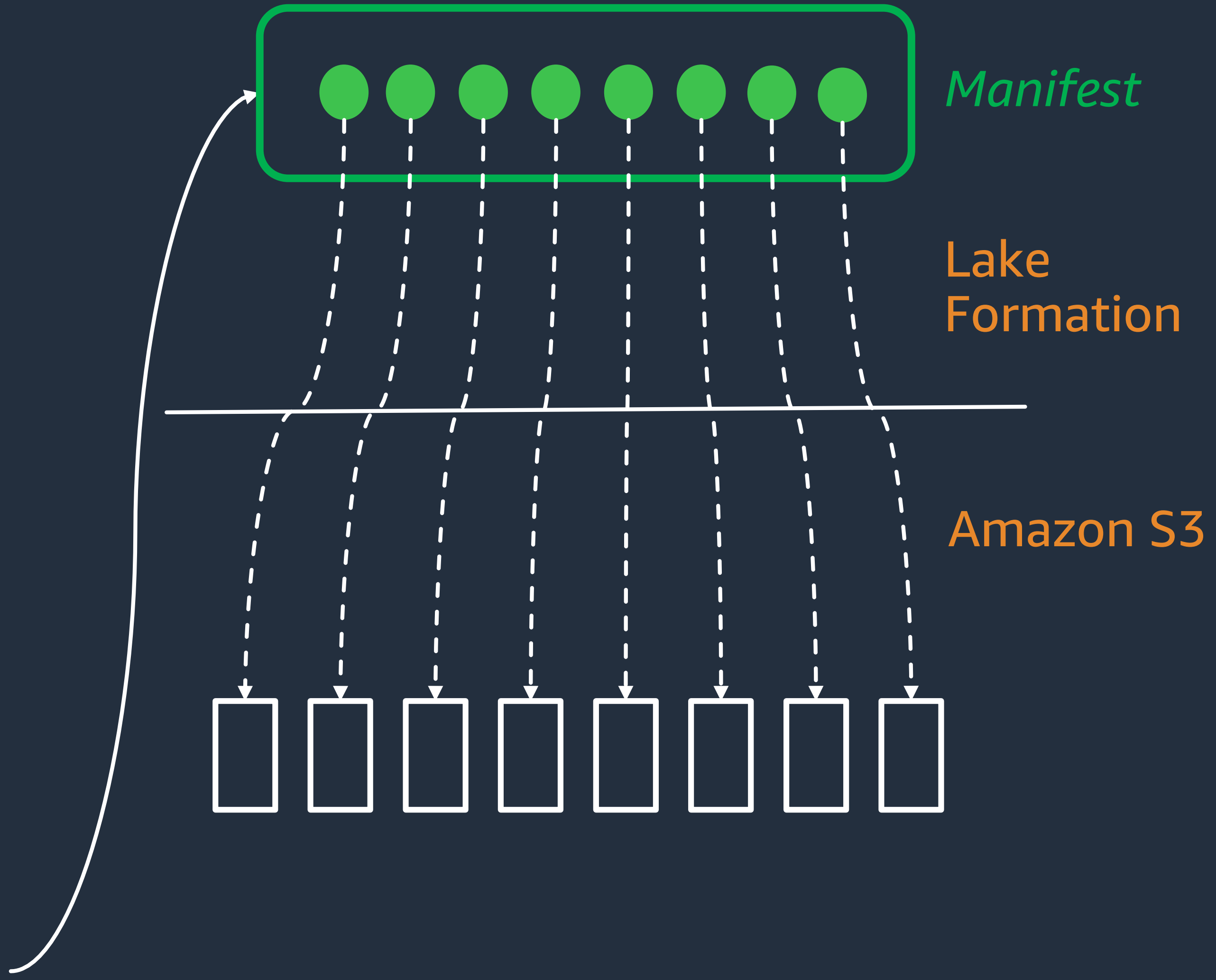
PREVIEW

Apache Hive-style tables



S3://IOTDeviceData/region=Americas/year=2020/month=11/day=30/data1.parquet
S3://IOTDeviceData/region=Americas/year=2020/month=11/day=30/data2.parquet
S3://IOTDeviceData/region=Americas/year=2020/month=11/day=30/data3.parquet
...

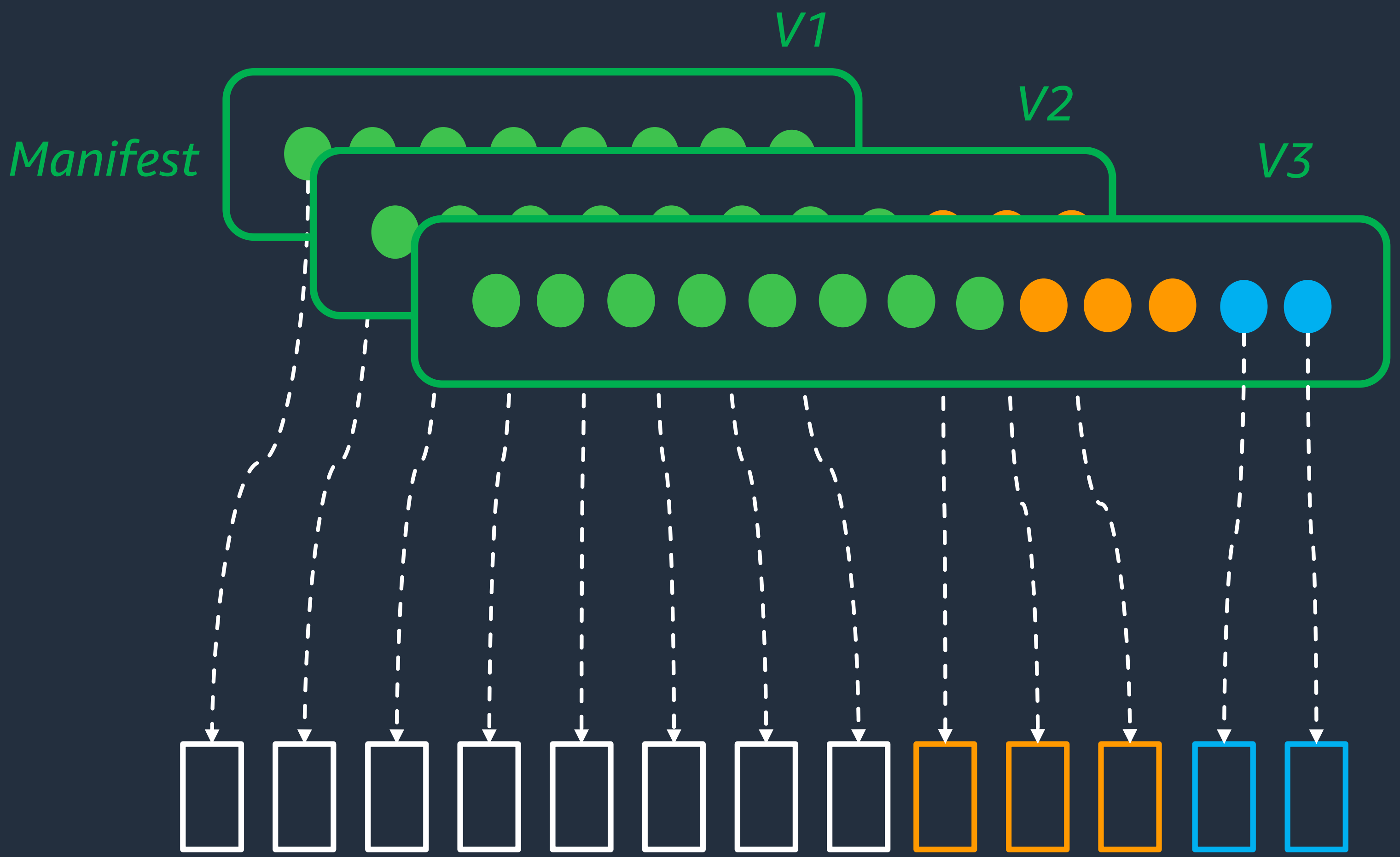
Governed Table



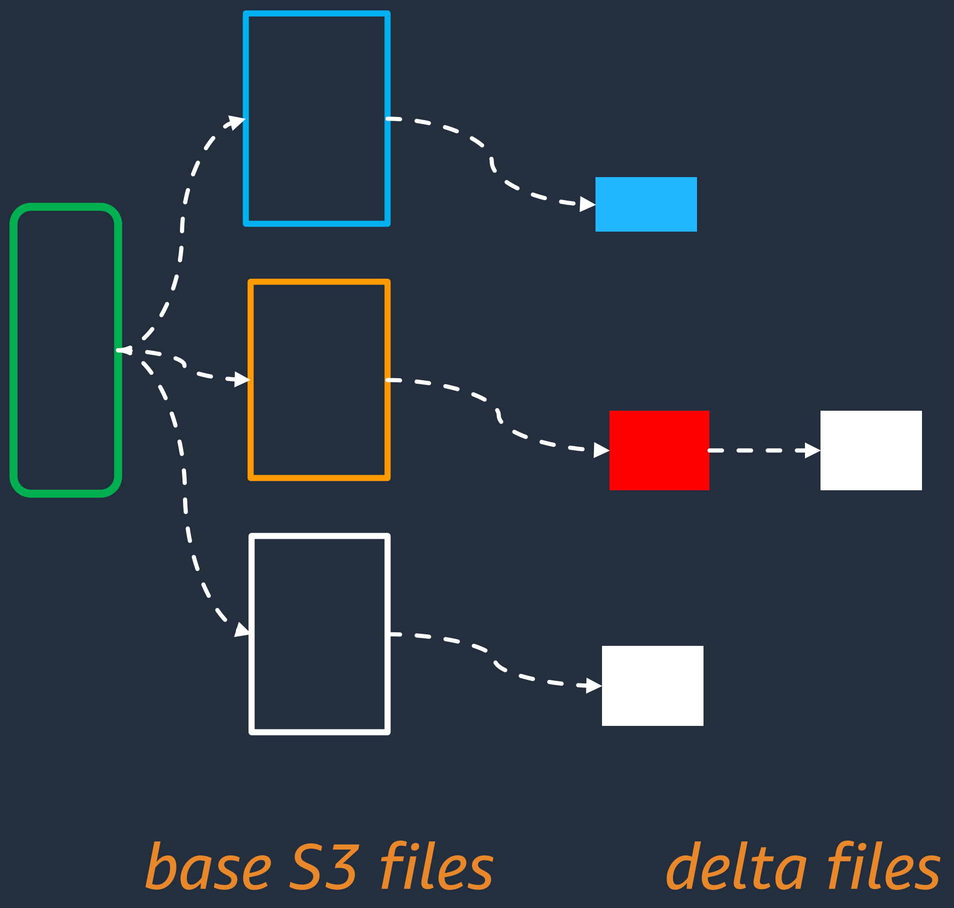
Governed Tables: Transactions

PREVIEW

Table-based txn API:
Add or remove files from manifest



Row-based txn API:
Insert, update, delete rows
delta files contain edits



Transactions make data lakes trustworthy

hudl

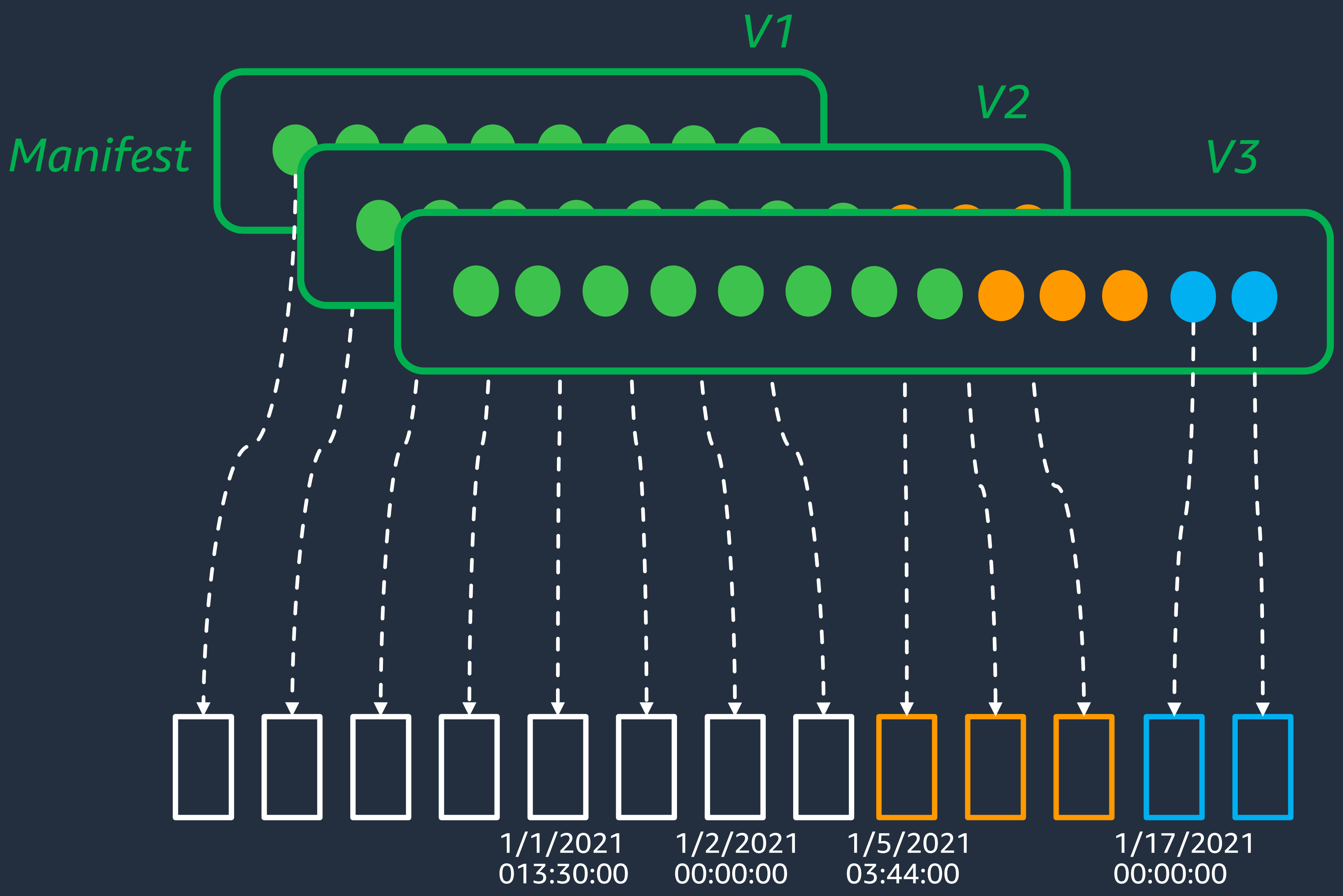
“. . . Transactional ETL processes are an important part of how we ensure data integrity and . . . required additional development time and complexity. We’re excited about AWS Lake Formation Transactions’ ability to simplify our ETL and reduce the overall effort needed to produce trustworthy data in our data lake."

Rob Hruska
Engineering Director
Hudl

aws

Governed Tables: Time travel

PREVIEW



```
SELECT * FROM gov_table
FOR SYSTEM_VERSION AS OF V2
```

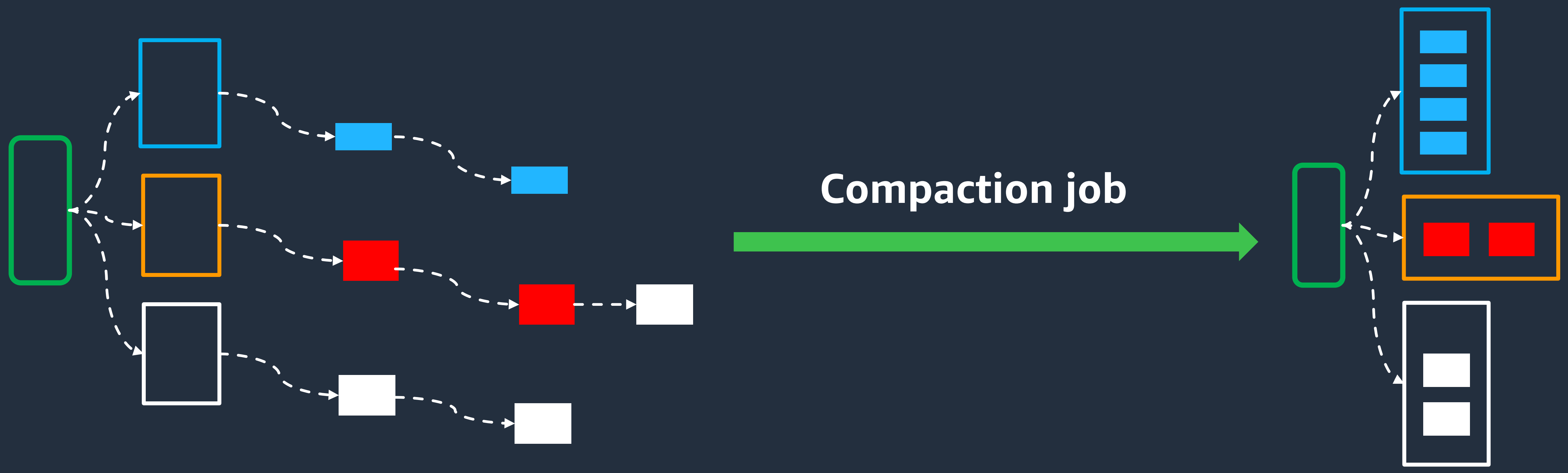
```
SELECT * FROM gov_table
FOR SYSTEM_TIME AS OF
‘2021-01-05 00:00:00’
```

Governed Tables: Storage optimizer

PREVIEW

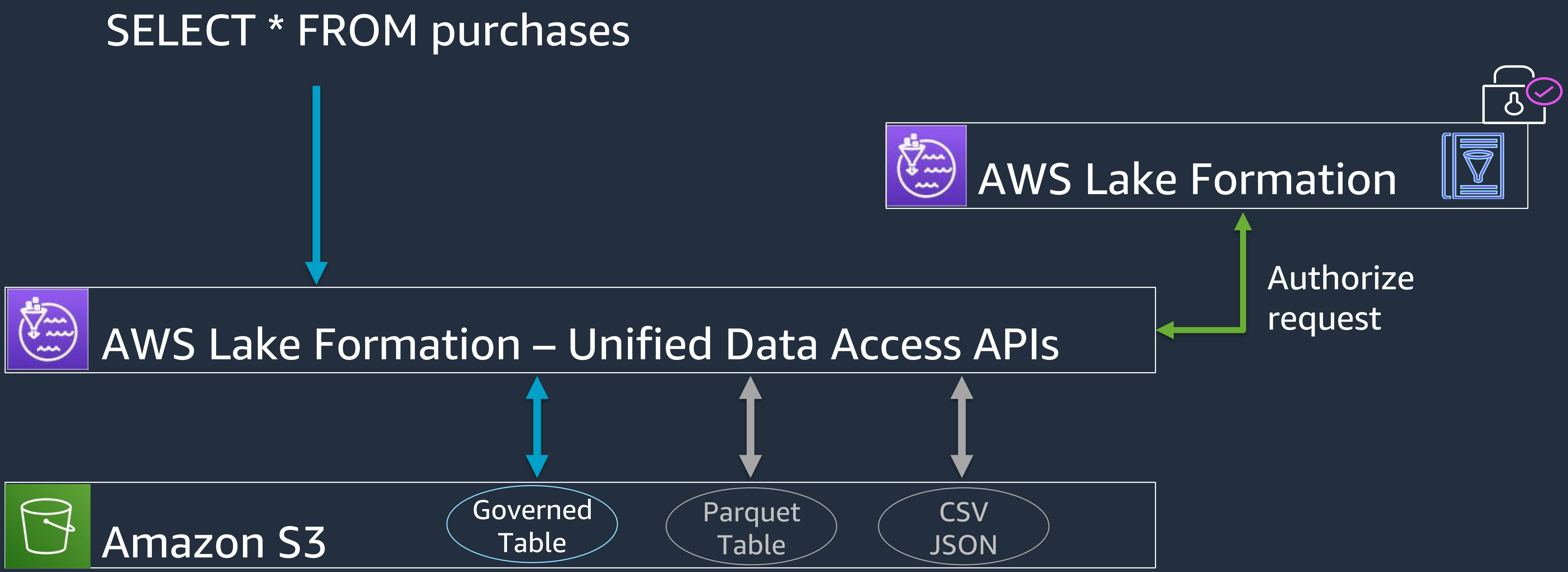
Automatic Compaction

Continuously monitors changes to underlying data of governed tables



New way to securely access data

PREVIEW



SELECT * FROM purchases

AWS Lake Formation

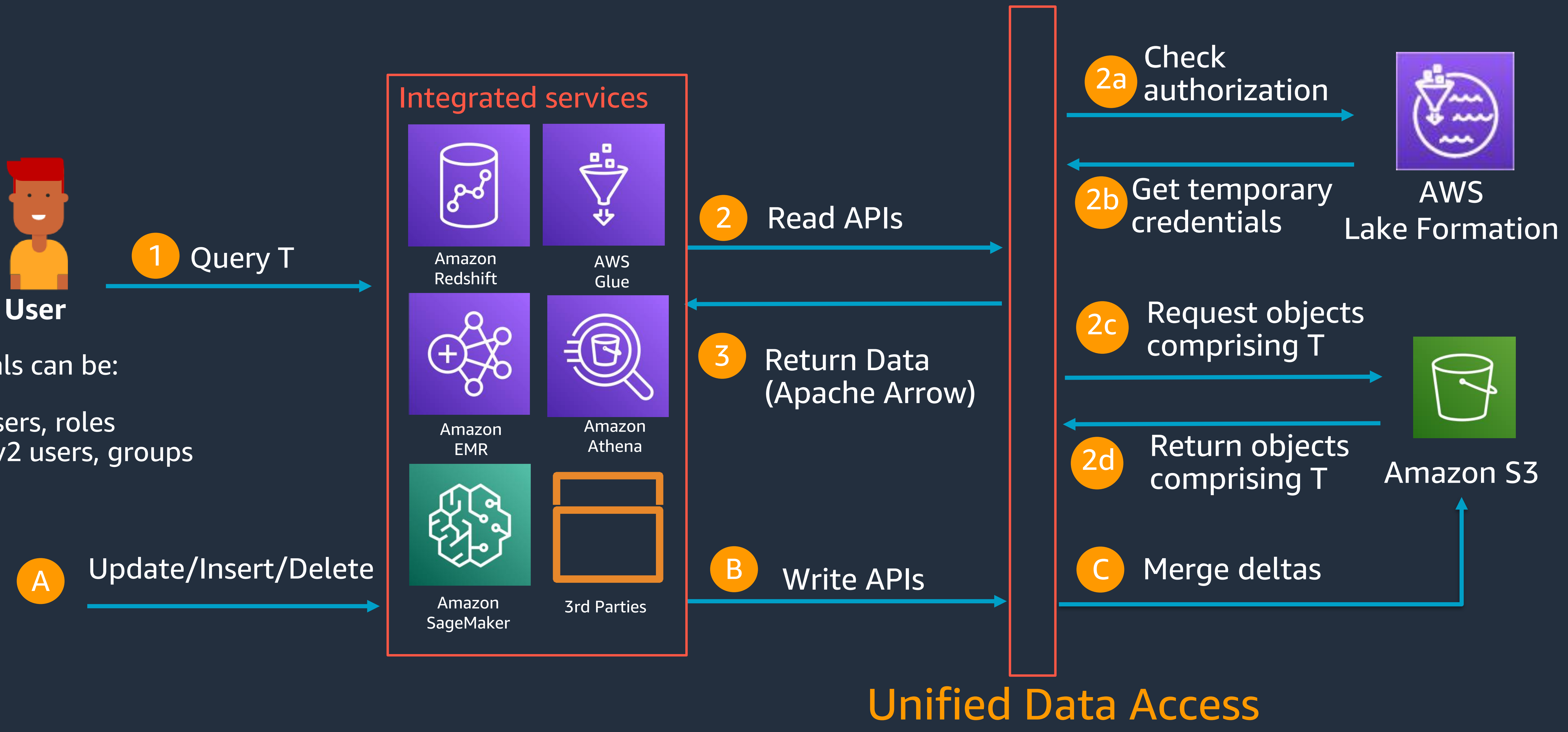
Authorize request

AWS Lake Formation – Unified Data Access APIs

Amazon S3

AWS Lake Formation securely manages reads and writes

PREVIEW



Principals can be:

- IAM users, roles
- SAMLv2 users, groups

Challenges in managing security and governance

VISIBILITY & OWNERSHIP

DATA SPRAWL
MINIMAL VISIBILITY
DIFFICULT TO SECURE

ENFORCING PERMISSIONS

GAPS IN CAPABILITIES
INCONSISTENT ENFORCEMENT

DATA DUPLICATION

REQUIRE ETL
MULTIPLE COPIES
COSTLY

Central governance

CENTRAL CATALOG

VIEW & MANAGE ASSETS
DATA CLASSIFICATION
ACCESS POLICIES
AUDIT

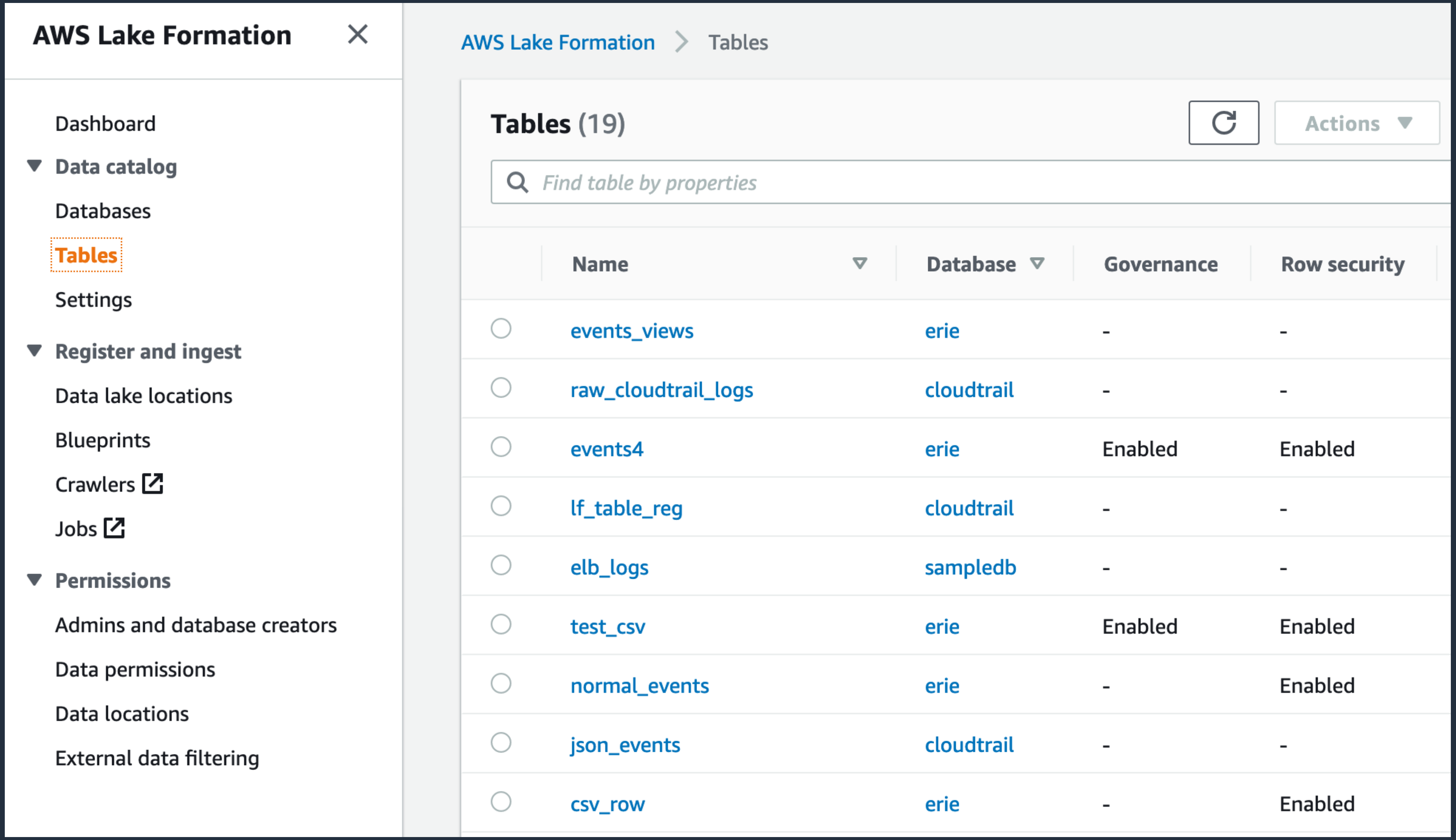
FINE GRAIN PERMISSIONS

CENTRALLY ENFORCED
TABLE | COLUMN | ROW
ATTRIBUTE

DATA SHARING

SEAMLESS SHARING
UBIQUITOUS
TOOL AGNOSTIC

AWS Lake Formation central catalog

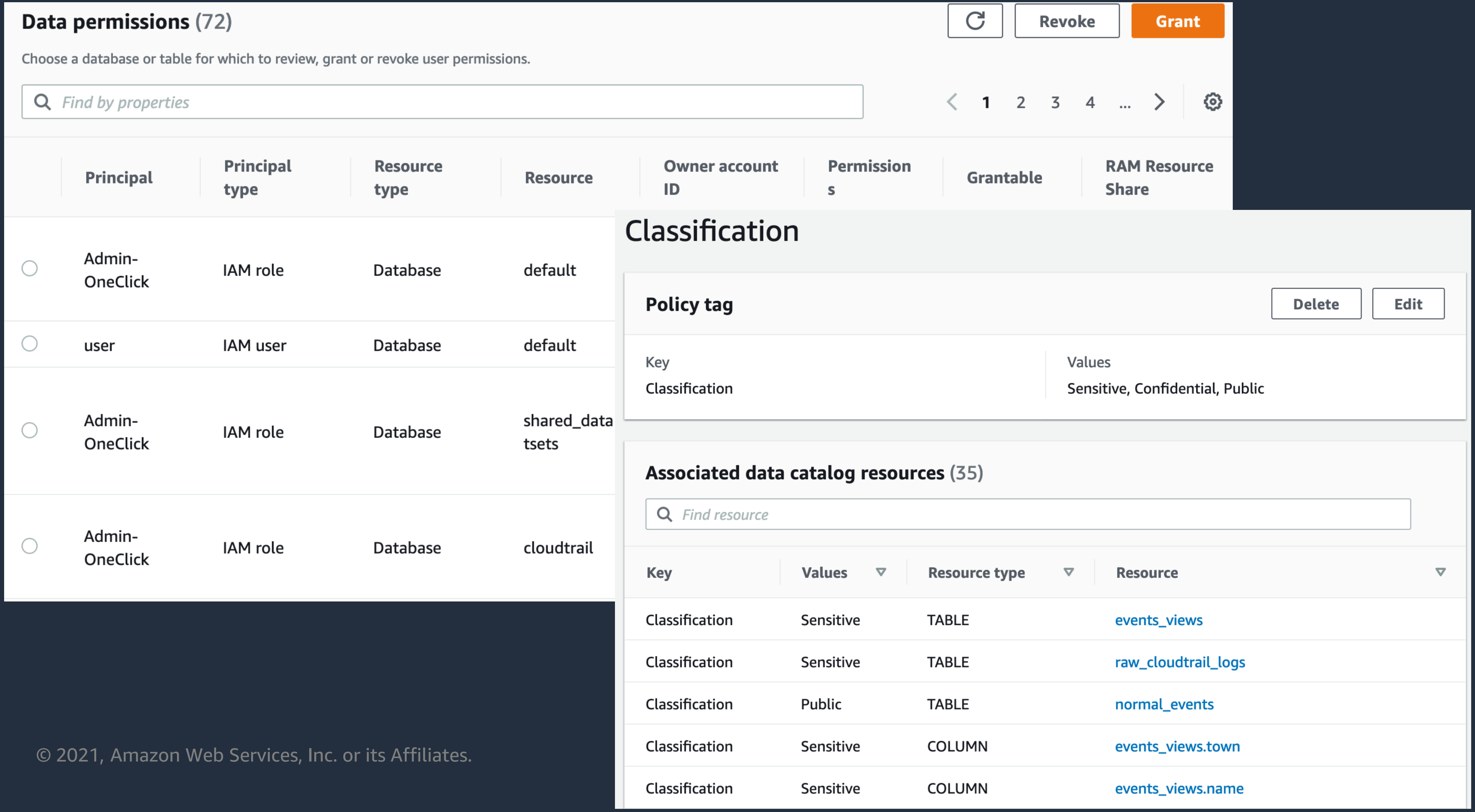


Tables (19)

	Name	Database	Governance	Row security
○	events_views	erie	-	-
○	raw_cloudtrail_logs	cloudtrail	-	-
○	events4	erie	Enabled	Enabled
○	lf_table_reg	cloudtrail	-	-
○	elb_logs	sampledb	-	-
○	test_csv	erie	Enabled	Enabled
○	normal_events	erie	-	Enabled
○	json_events	cloudtrail	-	-
○	csv_row	erie	-	Enabled

Data analysts & scientists

AWS Lake Formation central catalog



Data permissions (72)

Choose a database or table for which to review, grant or revoke user permissions.

	Principal	Principal type	Resource type	Resource	Owner account ID	Permissions	Grantable	RAM Resource Share
○	Admin-OneClick	IAM role	Database	default				
○	user	IAM user	Database	default				
○	Admin-OneClick	IAM role	Database	shared_datatsets				
○	Admin-OneClick	IAM role	Database	cloudtrail				

Classification

Policy tag

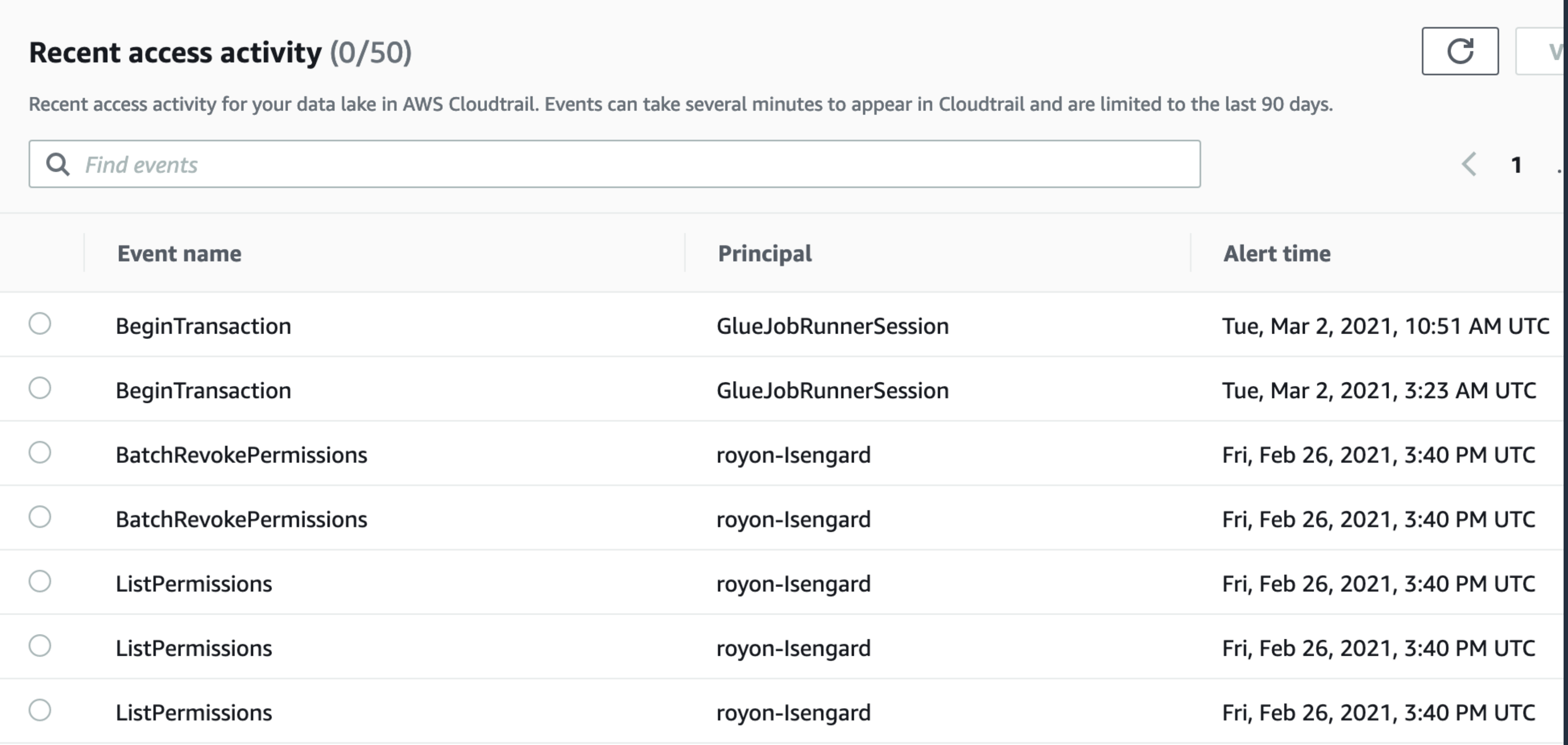
Key	Values
Classification	Sensitive, Confidential, Public

Associated data catalog resources (35)

Key	Values	Resource type	Resource
Classification	Sensitive	TABLE	events_views
Classification	Sensitive	TABLE	raw_cloudtrail_logs
Classification	Public	TABLE	normal_events
Classification	Sensitive	COLUMN	events_views.town
Classification	Sensitive	COLUMN	events_views.name

Data steward

AWS Lake Formation central catalog



Recent access activity (0/50)

Recent access activity for your data lake in AWS Cloudtrail. Events can take several minutes to appear in Cloudtrail and are limited to the last 90 days.

Find events

	Event name	Principal	Alert time
○	BeginTransaction	GlueJobRunnerSession	Tue, Mar 2, 2021, 10:51 AM UTC
○	BeginTransaction	GlueJobRunnerSession	Tue, Mar 2, 2021, 3:23 AM UTC
○	BatchRevokePermissions	royon-Isengard	Fri, Feb 26, 2021, 3:40 PM UTC
○	BatchRevokePermissions	royon-Isengard	Fri, Feb 26, 2021, 3:40 PM UTC
○	ListPermissions	royon-Isengard	Fri, Feb 26, 2021, 3:40 PM UTC
○	ListPermissions	royon-Isengard	Fri, Feb 26, 2021, 3:40 PM UTC
○	ListPermissions	royon-Isengard	Fri, Feb 26, 2021, 3:40 PM UTC

Data administrator

NEW

Lake Formation: Tag-based access control

Simple to manage and scale permissions

Decouple permissions from resources

Tag ontology for annotation and classification

Lake Formation: Tag based access control

NEW

Easily scale permission enforcement

1. Define tag ontology centrally
2. Tag databases, table, columns
3. Grant principals access on tags

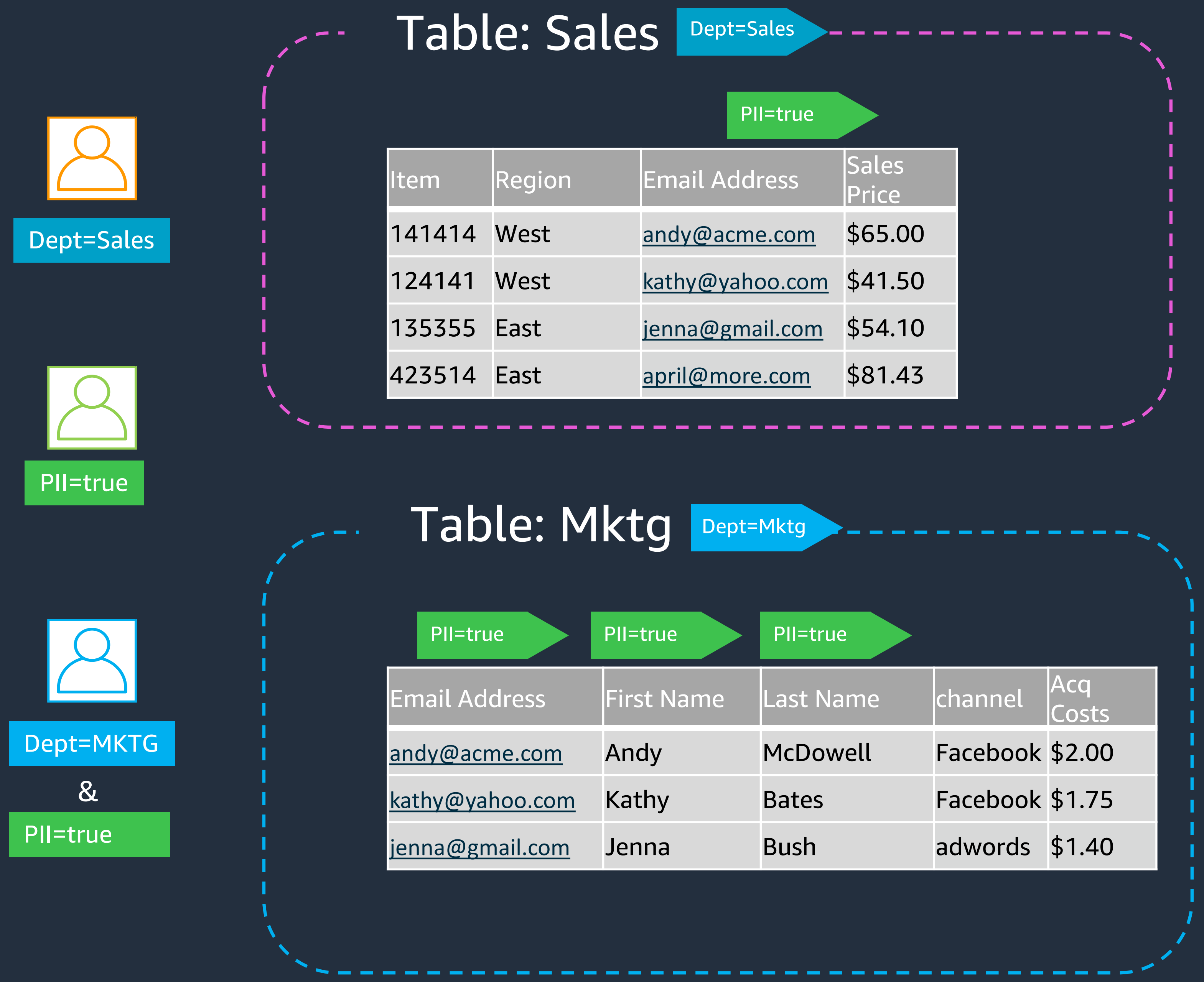


Table: Sales Dept=Sales

PII=true

Item	Region	Email Address	Sales Price
141414	West	andy@acme.com	$65.00
124141	West	kathy@yahoo.com	$41.50
135355	East	jenna@gmail.com	$54.10
423514	East	april@more.com	$81.43

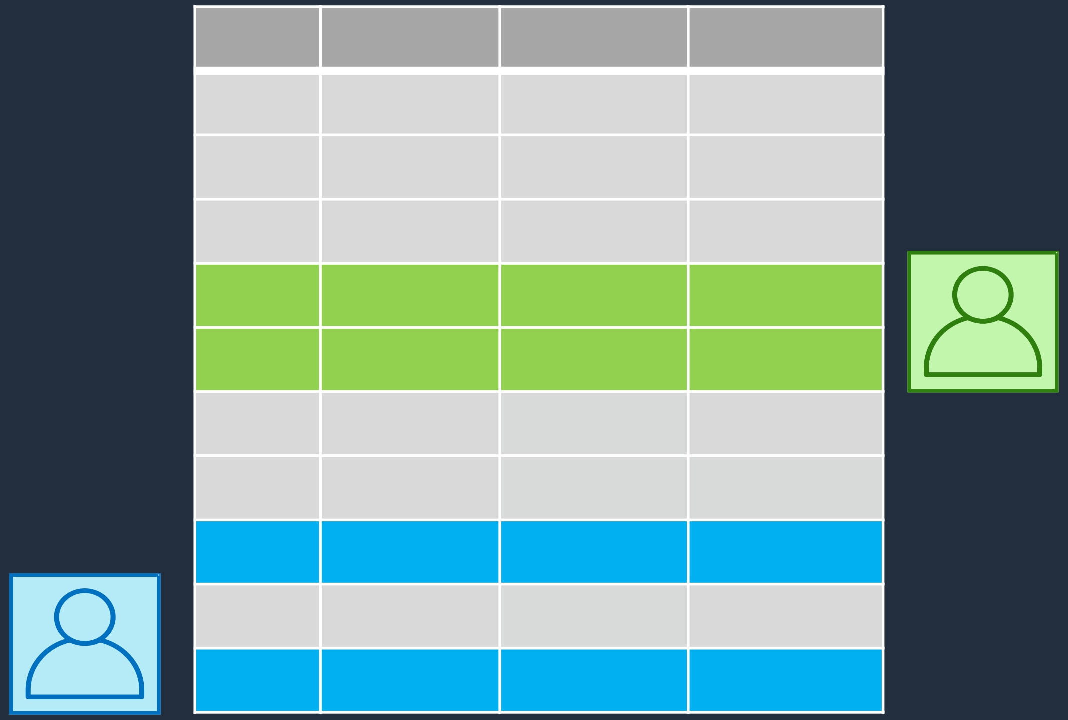
Table: Mktg Dept=Mktg

PII=true PII=true PII=true

Email Address	First Name	Last Name	channel	Acq Costs
andy@acme.com	Andy	McDowell	Facebook	$2.00
kathy@yahoo.com	Kathy	Bates	Facebook	$1.75
jenna@gmail.com	Jenna	Bush	adwords	$1.40

AWS Lake Formation row-level security

PREVIEW



Row-level security policies

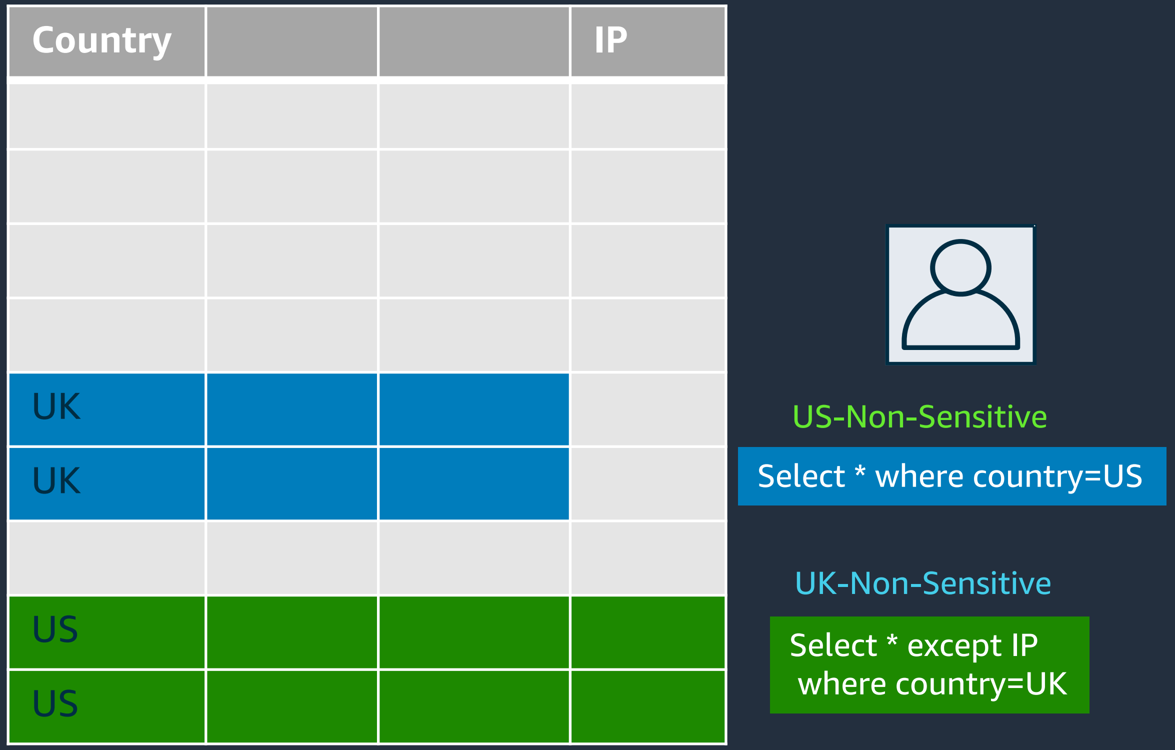
Row filter expressions are
"WHERE" clause in "PartiQL"

Supports many S3-based table formats
Open and managed
Governed, Amazon Redshift data shares, Apache Hive
Apache Iceberg, Apachi Hudi, Delta Lake, . . .

Easy to audit permissions and access

AWS Lake Formation cell-level security

PREVIEW



Country			IP
UK			
UK			
US			
US			

US-Non-Sensitive

Select * where country=US

UK-Non-Sensitive

Select * except IP
where country=UK

Country			IP
UK			********
UK			********
US			
US			

Effective access with masked IP column

Cell-level policies extends row-level security

Define cell regions by selecting columns and rows

Included columns: name, age, country
Row filter: country = ‘US’

Automatically mask excluded cells

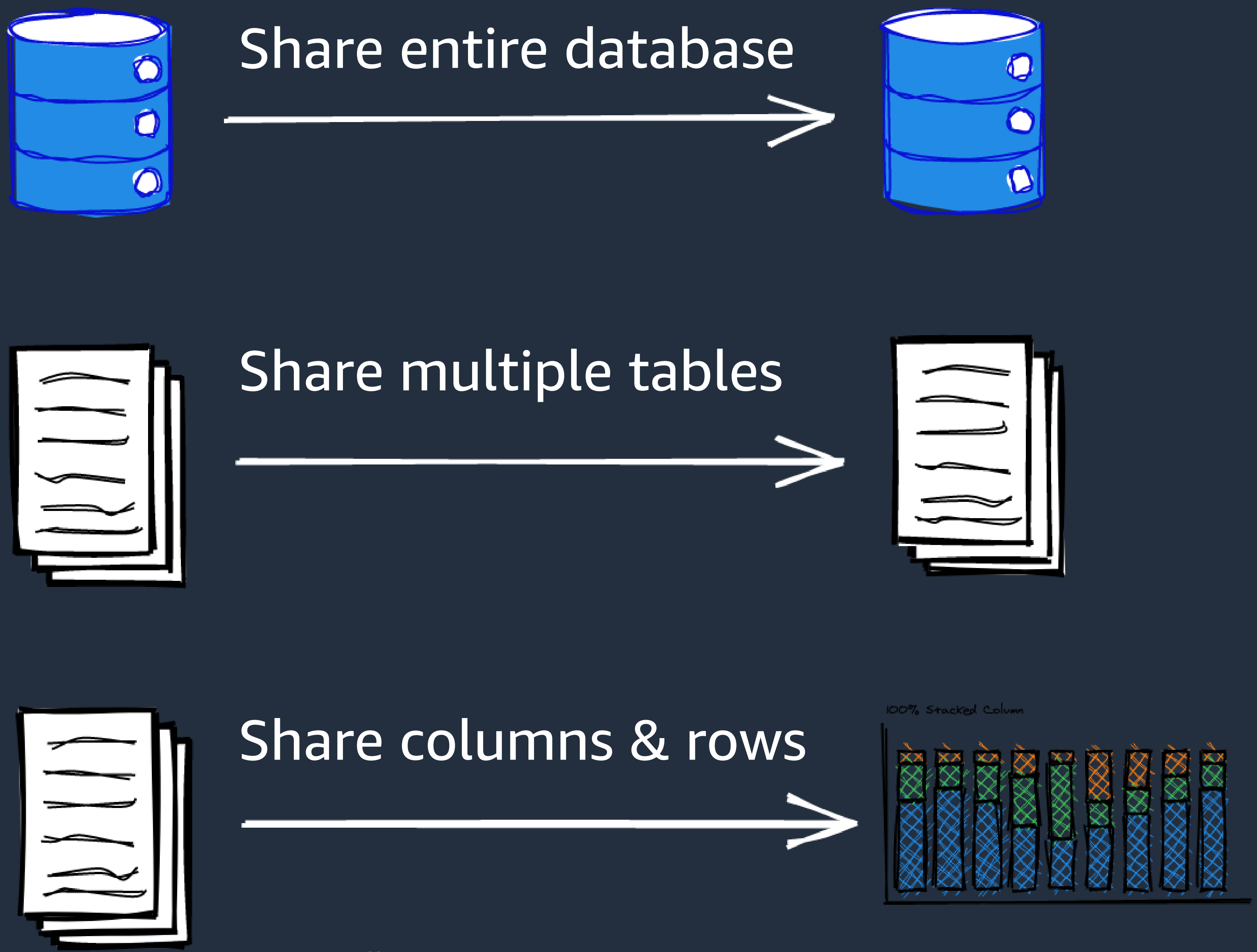
Easy to audit permissions and access

Managing data lake permissions

“We found AWS Lake Formation easy to use to build and secure our data lake. Without AWS Lake Formation, we would have to make constant access policy updates to Amazon S3 when we added more users and data . . . With the adoption of AWS Lake Formation, we are able to . . . reduce Amazon S3 policy edits by over 90% . . ."

Hisatoshi Imaoka
Tech. Lead Data Infrastructure
freee K.K.

Data sharing made simple with Lake Formation



Share entire database

Share multiple tables

Share columns & rows

Amazon QuickSight

Amazon Redshift

AWS Glue

Amazon SageMaker

Amazon Athena

3rd Party ISV

Amazon EMR

Sharing data lakes across accounts

GoDaddy

“. . . We are building a hub-and-spoke architecture using AWS Lake Formation where data producers can publish their sharable data to a centralized data catalog and data subscribers can request access to that data from the centralized data catalog."

Charles Beadnall
CTO
GoDaddy

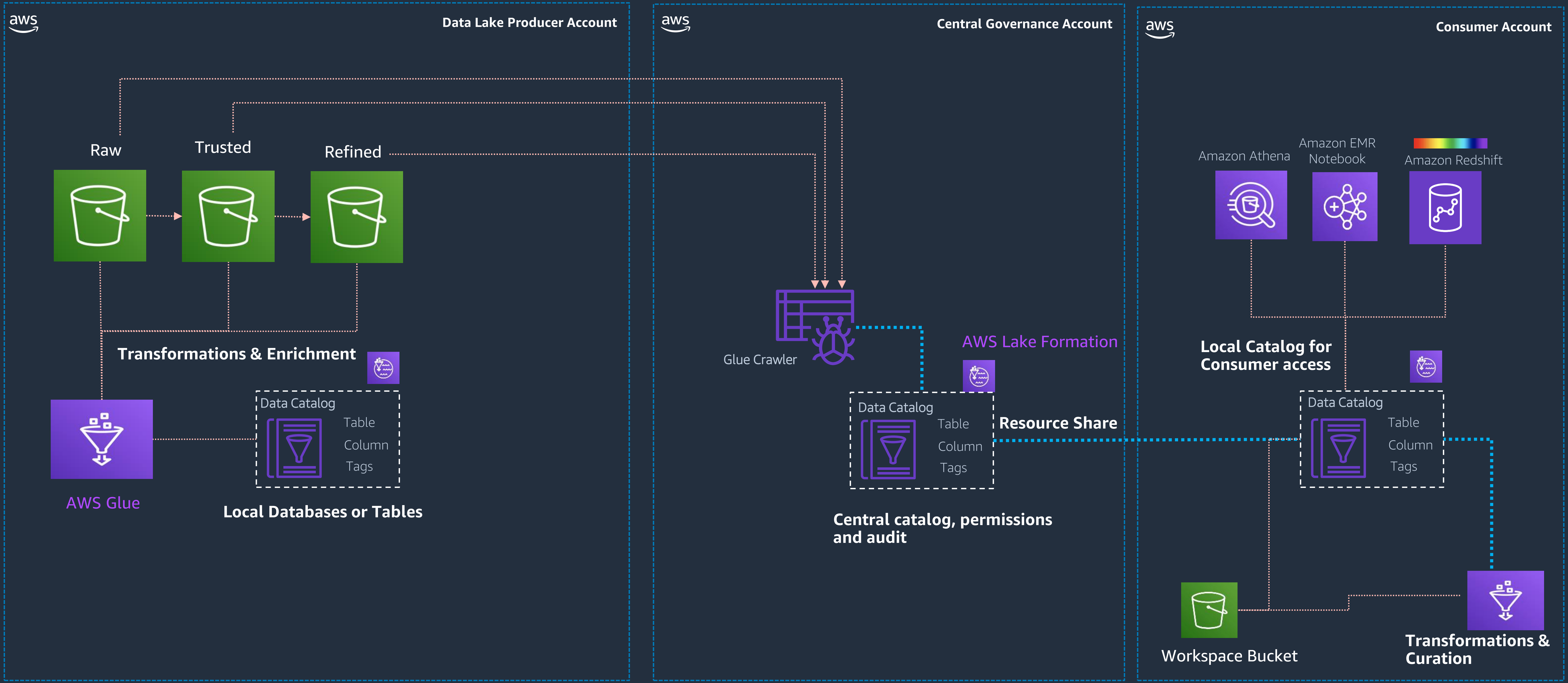
D2C

“. . . To work on the data lake with multiple AWS accounts, we need a central metadata catalog . . . Using the cross account database/table sharing of AWS Lake Formation, we are able to achieve our goal easily, without affecting the existing workload . . . the ability to manage data by column enables more sophisticated data management."

Rinichiro Nagatomo
Technical lead
Data management platform
D2C Inc.

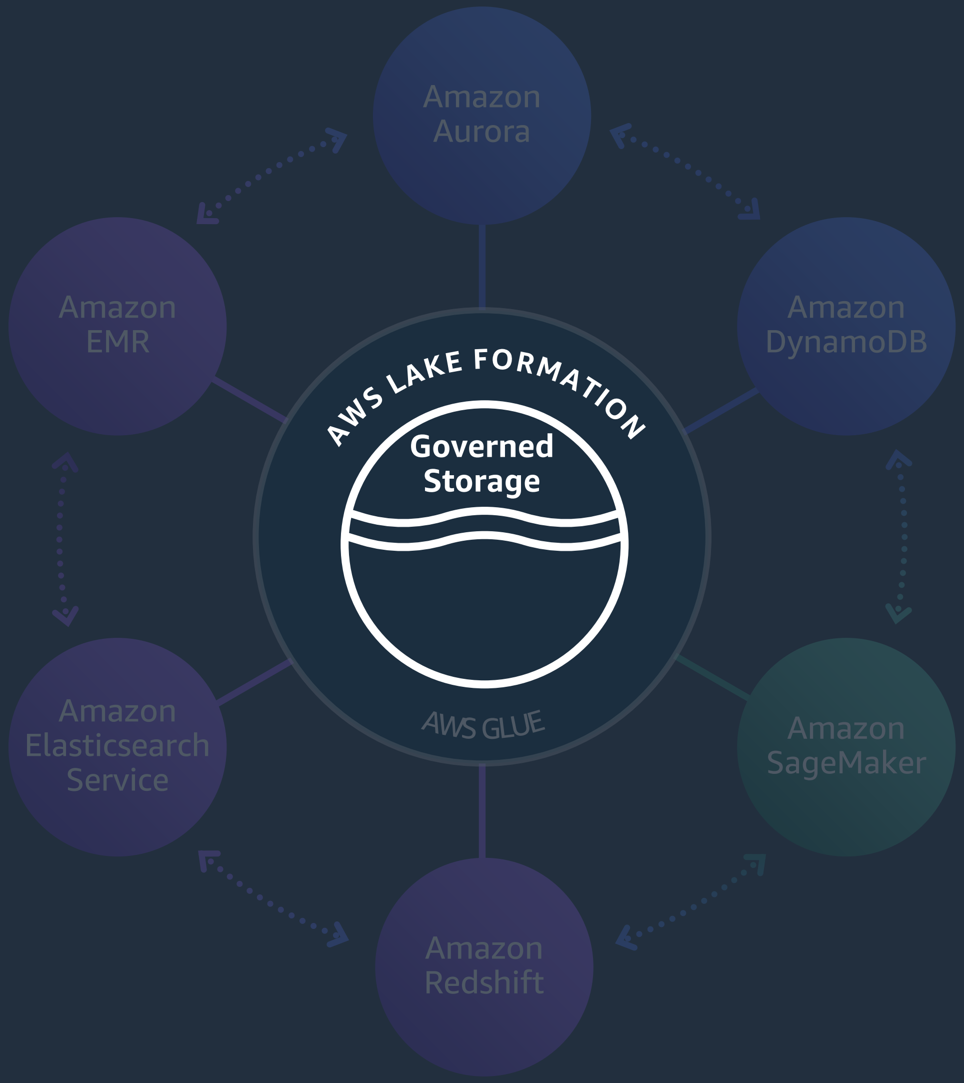
aws

Data Mesh with Central Governance



AWS Lake Formation

Build a secure data lake in days



Build data lakes quickly

Move, store, update, and catalog your data faster
Automatically organize and optimize your data

Simplify security management

Centrally define and enforce security, governance, and auditing policies

Easily discover and share data

Catalog all of your data assets and easily share datasets between consumers

Getting started

Lake Formation is generally available in the following Regions

Americas	EMEA	Asia
Montreal	Frankfurt	Mumbai
North California	Ireland	Seoul
Northern Virginia	London	Singapore
Ohio	Milan	Sydney
Oregon	Stockholm	Tokyo
Sao Paulo	Paris	Hong Kong
GovCloud		Beijing

Start today: https://aws.amazon.com/lake-formation

Sign up for the preview: https://pages.awscloud.com/Lake_Formation_Feature_Preview.html

Thank you!

Roy Hasson

LinkedIn: /in/royhasson